NOTES AND CORRESPONDENCE

A Scoring System for Probability Forecasts of Ranked Categories[1]

EDWARD S. EPSTEIN

Dept. of Meteorology & Oceanography, The University of Michigan, Ann Arbor[2]

26 May 1969

1. Introduction

Consider probability forecasts of the four temperature classes: $T\leqslant 0\text{F}$, $0\text{F}<T\leqslant 20\text{F}$, $20\text{F}<T\leqslant 40\text{F}$, $T>40\text{F}$. If two forecasts were (0.1,0.3,0.5,0.1) and (0.5,0.3,0.1,0.1) and the last category, $T>40\text{F}$, were observed, all proper scoring systems now in use would assign the same score to both forecasts. Most would agree, though, that the former was a somewhat better forecast. This conclusion is based on the notion that categories 3 and 4 are "closer" to one another than categories 1 and 4. The concept of "distance" in this sense does not exist in any of the proper scoring systems previously proposed. For example, in the geometrical framework which is so natural for representing the "probability score" (Epstein and Murphy, 1965), the vertices representing the various weather states are always equidistant from all other vertices.

In this note a new scoring rule is presented in which the ranking of the several alternative weather states is implicit. This rule is derived as an extension of methods first presented by Murphy (1966). One of the particular advantages of this method is that one is assured that the resulting scoring rule is proper[3] (Murphy and Epstein, 1967).

It should be mentioned, as a note of caution, that although the derivation given below is based on the concepts of decision theory and utility, the specific assumptions are too artificial to permit identifying the resulting score with the value of the forecast. Values of forecasts can only be assessed for specific users and their specific utility matrices. While the scoring rule given below may be a measure of value under sufficiently unusual circumstances, it is best to think of it only as a useful proper scoring rule for ranked categories that can serve as a convenient standard.

2. Derivation

Consider a decision situation in which there are K possible weather states and K possible actions. The possible actions A_i, $i=1,\ .\ .\ .\ ,K$, are to take successively less complete protective measures. The costs of these different degrees of protection are $C(K-i)/(K-1)$ where C is the cost of complete protection. The possible weather states, W_j, $j=1,\ .\ .\ .,K$, range from that requiring no protection ($j=K$) to that for which full protection is required ($j=1$). If $j\geqslant i$, the protection is adequate and no further costs are incurred. If, however, $j<i$, the weather creates additional losses in the amounts $L(i-j)/(K-1)$, L being the loss which occurs when no protective action is taken and the most severe possible weather occurs. The total "cost" of taking action A_i when weather W_j subsequently occurs is then

$$c_{ij}=\begin{cases}C(K-i)/(K-1), & i\leq j\\ [C(K-i)+L(i-j)]/(K-1), & i>j\end{cases}$$

This is a natural extension to K categories of the usual 2×2 cost-loss matrix.

It is convenient to express the decision matrix in a standardized form such that the most preferred outcome has a "value" of $+1$, while the least desired result has the "value" 0. For this purpose, then, we define $u_{ij}=1-c_{ij}/L$. The u_{ij}, are now treated as though they were the elements of a utility matrix.

The decision rule corresponding to this matrix depends only on the ratio C/L and the forecast $(p_1,\ .\ .\ .,p_k)$, which we assume are the decision-maker's

[1] Contribution No. 162 from the Department of Meteorology & Oceanography, The University of Michigan.

[2] The research reported here was carried out while the author was a visiting scientist at the Institute of Meteorology, University of Stockholm, Sweden.

[3] Although there exist heuristic arguments requiring this statement to be true, no general proof is known to the author. Murphy (1969b) has shown it to be valid for the 2×2 cost-loss decision situation, and his proof may be extended to cover the decision situation discussed here.

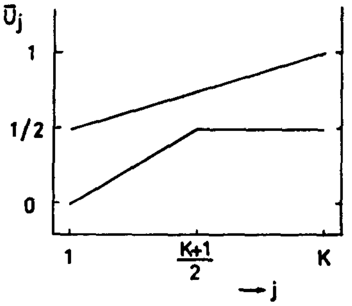

FIG. 1. Bounds on the expected kernal utility of a prediction.

probabilities as well as the forecaster's. Specifically, to maximize expected utility, the decision maker would take action k whenever

$$\sum_{j=1}^{k-1} p_j < C/L < \sum_{j=1}^{k} p_j.$$

The utility the decision maker actually achieves [the "kernal utility of a prediction," in the terminology of Murphy (1966)] depends on which weather state actually occurs. If this happens to be W_j, then this may be written as

$$U_j = \sum_{i=1}^{K} u_{ij} d_i(\mathbf{p}, C/L),$$

where

$$d_i(\mathbf{p}, C/L) = \begin{cases} 1, & \text{if } \sum_{j=1}^{i-1} p_j < C/L \leq \sum_{j=1}^{i} p_j \\ 0, & \text{otherwise} \end{cases}$$

Treating C/L as a random variable, with a density $f(C/L)$ allows us to define an average, or expected kernal utility of a prediction, as

$$\bar{U}_j = \int_0^1 \sum_{i=1}^{K} u_{ij} d_i(\mathbf{p}, X) f(X) dX.$$

Making the appropriate substitution for u_{ij}, this becomes

$$\bar{U}_j = \frac{1}{K-1} \int_0^1 \left\{ \sum_{i=1}^{j} [K-1-X(K-i)] d_i(\mathbf{p},X) + \sum_{i=j+1}^{K} [j-i+K-1-X(K-i)] d_i(\mathbf{p},X) \right\} f(X) dX,$$

$$= \frac{1}{K-1} \left\{ \int_0^1 \sum_{i=1}^{K} [K-1-X(K-i)] d_i(\mathbf{p},X) f(X) dX + \int_0^1 \sum_{i=j+1}^{K} (j-i) d_i(\mathbf{p},X) f(X) dX \right\},$$

$$= \frac{1}{K-1} \left\{ \sum_{i=1}^{K} \int_{l_{i-1}}^{l_i} [K-1-X(K-i)] f(X) dX + \sum_{i=j+1}^{K} \int_{l_{i-1}}^{l_i} (j-i) f(X) dX \right\},$$

where $l_i = \sum_{n=1}^{i} p_n$ are the limits of integration imposed by the step function $d_i(\mathbf{p},X)$.

The density $f(X)$ is now, arbitrarily, taken to be the uniform distribution, $f(X)=1$, $0 \leqslant X \leqslant 1$. This gives a particularly simple result, and is the same assumption used by Murphy (1966). Murphy (1969a) has shown how appropriate choices for $f(X)$ can lead to a family of scoring rules in the unranked situation. By the same token the introduction here of a family of densities for $f(X)$ would lead to a family of scoring rules for ranked categories. For the present there seems to be no significant benefit from such an exercise. The uniform distribution gives

$$\bar{U}_j = 1 - \frac{1}{2(K-1)} \sum_{i=1}^{K} (K-i) \left[\left(\sum_{n=1}^{i} p_n\right)^2 - \left(\sum_{n=1}^{i-1} p_n\right)^2 \right] + \frac{1}{K-1} \sum_{i=j+1}^{K} (j-i) p_i,$$

$$= 1 - \frac{1}{2(K-1)} \sum_{i=1}^{K-1} \left(\sum_{n=1}^{i} p_n\right)^2 - \frac{1}{K-1} \sum_{i=j+1}^{K} (i-j) p_i.$$

The expected kernal utility has a maximum value of $\frac{1}{2}(K+j-2)/(K-1)$; better results are possible when "better" weather occurs. The minimum value is $\frac{1}{2}$ for $j \geqslant (K+1)/2$ and is $(j-1)/(K-1)$ for $j \leqslant (K+1)/2$. These facets of the result are illustrated in Fig. 1. Which score, within the indicated range, is achieved, depends of course on how "good" the forecast is.

In order to produce a scoring scheme which is less dependent for the absolute value of the score on what weather subsequently occurs, and more dependent on the "goodness" of the forecast, we have modified $\bar{U}_j$ by adding a similar score calculated as though the category $j=K$ were the most severe, and then subtracting $\frac{1}{2}$ from

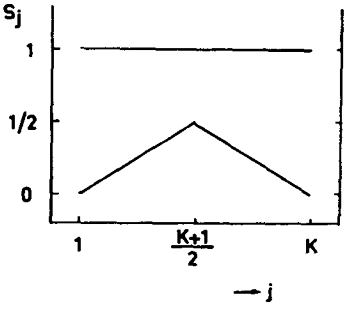

FIG. 2. Bounds on S_j, the "ranked probability score."

the result. This gives our recommended scoring rule

$$S_j=\frac{3}{2}-\frac{1}{2(K-1)}\sum_{i=1}^{K-1}\Big[\Big(\sum_{n=1}^{i}p_n\Big)^2+\Big(\sum_{n=i+1}^{K}p_n\Big)^2\Big]$$

$$-\frac{1}{K-1}\sum_{i=1}^{K}|i-j|p_i.$$

The range of possible values of S_j is shown, as a function of j, in Fig. 2.

3. Discussion

A perfect forecast ($p_j=1$ and W_j occurs) always results in a score of 1. The worst possible forecast ($p_1=1$ and W_K occurs, or *vice versa*) receives a score of 0. If the weather that occurs is in the middle of the range of possible values, no forecast can be as bad as that and the minimum scores are therefore larger.

If the scoring rule S_j is applied to the two forecasts mentioned in Section 1, one obtains the scores shown in Table 1. If the fourth category occurs, the first forecast receives a substantially better score.

It may also be noted, in Table 1, that given the forecast (0.5,0.3,0.1,0.1) one obtains the same value for S_j whether category 1 or 2 is observed. This is an example of the characteristic of this scoring rule to "consider" both the categories to which the bulk of the probability is assigned, and also the "expected" category implied by the distribution of probabilities. This is further illustrated in Table 2. A forecast of $\frac{1}{2}$ for both p_1 and p_K always results in a score of 0.75, whatever the total number of categories and whichever event occurs. The more central the category that occurs, the "further" one is from the categories to which probabilities are assigned, but the "closer" one is to the "expected" category.

The case where the forecast is $1/K$ for each category may be of special interest and can be treated algebraically. The ranked probability score for this case reduces to

$$S_j^*=\frac{2}{3}+\frac{1}{6K}+\frac{(K-j)(j-1)}{K(K-1)}.$$

If one of the extreme events ($j=1$ or $j=K$) occurs, this score will take on a low value of $(4K+1)/6K$. The highest possible score given this special forecast is $(11K-1)/12K$, corresponding to $j=\frac{1}{2}(K+1)$, K odd. If K is even the maximum score is $11/12-(K+2)/[12K(K-1)]$. Thus, $\frac{2}{3}<S_j^*<11/12$ for all K.

TABLE 1. Ranked probability scores for two illustrative forecasts.

Observed category	Forecast (0.1,0.3,0.5,0.1)	 (0.5,0.3,0.1,0.1)
1	0.61	0.90
2	0.87	0.90
3	0.94	0.70
4	0.67	0.43

TABLE 2. Ranked probability scores for some "standard" forecasts ($K=6$).

	Observed weather category					
Forecast	1	2	3	4	5	6
(1,0,0,0,0,0)	1.00	0.80	0.60	0.40	0.20	0.00
(0,1,0,0,0,0)	0.80	1.00	0.80	0.60	0.40	0.20
(0,0,1,0,0,0)	0.60	0.80	1.00	0.80	0.60	0.40
$(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6})$	0.69	0.83	0.89	0.89	0.83	0.69
$(\frac{1}{2},0,0,0,0,\frac{1}{2})$	0.75	0.75	0.75	0.75	0.75	0.75
$(\frac{1}{2},\frac{1}{2},0,0,0,0)$	0.95	0.95	0.75	0.55	0.35	0.15
$(0,0,\frac{1}{2},\frac{1}{2},0,0)$	0.55	0.75	0.95	0.95	0.75	0.55
$(0,\frac{1}{6},\frac{1}{3},\frac{1}{3},\frac{1}{6},0)$	0.61	0.81	0.92	0.92	0.81	0.61
$(\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0,0)$	0.89	0.96	0.89	0.69	0.49	0.29

Let us consider, in particular, the situation where the categories have been chosen such that their climatological relative frequencies are all equal. Then the forecast of $1/K$ as the probability of each cateogry is a climatological forecast, and the long term average score for such a forecast is $\sum_{j=1}^{k}S_j^*/K$. This quantity has the value $(5K-1)/6K$ which may, under appropriate conditions, serve as a base to determine a crude index of the incremental value, over climatology, of a set of forecasts.

The scoring rule S_j is a proper scoring system. This is assured by the manner of its derivation as an expectation of utility. The proof that S_j is proper, by the methods described by Murphy and Epstein (1967), is not difficult and is omitted here (Murphy, 1969c).

As a final comment, for the case $K=2$, S_j reduces to the familiar probability score. This of course it must, since in that case the derivation followed here becomes identical with that used by Murphy (1966) to derive the probability score from the elementary cost-loss matrix. The resemblence of the derivations justifies, I believe, the use of the term "ranked probability score" to refer to S_j.

REFERENCES

Epstein, E. S., and A. H. Murphy, 1965: A note on the attributes of probabilistic predictions and the probability score. *J. Appl. Meteor.*, **4**, 297–299.

Murphy, A. H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. Appl. Meteor.*, **5**, 534–537.

——, 1969a: Measures of the utility of probabilitic predictions in the cost-loss ratio decision situation in which knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteor.*, **8**, 863–873.

——, 1969b: On expected-utility measures in cost-loss ratio decision situations. *J. Appl. Meteor.*, **8**, 989–991.

——, 1969c: On the "Ranked Probability Score." *J. Appl. Meteor.*, **8**, 988–989.

——, and E. S. Epstein, 1967: A note on probability forecasts and "hedging." *J. Appl. Meteor.*, **6**, 1002–1004.